

# **Comparative Analysis for Oral Cleft Trio Data**

by

Jing Li

A dissertation submitted to Johns Hopkins University in conformity with the requirements for  
the degree of Master of Science in Biostatistics.

Baltimore, Maryland

April 2018

© Jing Li 2018 All rights reserved

## Abstract

The goal of this study is to compare data generated from two sequencing studies aimed at determining genetic causes of oral-facial clefts (OFCs) to assess whether the data are of similar quality and information content, and therefore could be combined to increase power for tests of genetic association. The purpose of this study is to find a reasonable approach to combine the two data sets to gain more statistical power for further studies. The first data set is from a previously published targeted sequencing (TS) study, which focused on 13 candidate genetic regions previously linked to or associated with risk to OFCs and included 1,409 case-parent trios of different population backgrounds, including 374 European trios, who we focus on here. The recently generated whole-genome sequencing (WGS) data was collected as part of the Gabriella Miller Kids First initiative, and contains 1,136 individuals (in approximately 378 case-parent trios) of European ancestry. We started by performing data cleaning of the WGS data based on the same quality control (QC) steps from the TS study, producing a clean data set of 981 individuals (in 327 trios). We then compared variant sets, and assessed concordance of genotype calls in individuals who were duplicated across the TS and WGS data sets ( $n=402$  in 134 trios). We then generated results from the genotypic transmission-disequilibrium test (gTDT) at common variants (i.e. those with minor allele frequency (MAF)  $\geq 0.01$ ), along with visualizations of patterns of linkage disequilibrium (LD) in these two data sets, with a focus on the region 8q24, which has previously been strongly associated with OFC among Europeans. Overall, good concordance and high similarity were observed in the sequence variants found in both data sets. We found combining the TS data and the WGS data provides increased power to detect association in the 8q24 region. Future work will be undertaken to use this combined data set and to perform more detailed comparative analysis across populations in this region.

Primary Reader: Margaret A. Taub

Secondary Reader: Terri H. Beaty

## **Acknowledgements**

I'd like to thank to Dr. Taub for advising me on this fantastic project and for helping me revising this dissertation.

I'd also like to thank to Dr. Beaty for the careful revision on this dissertation.

Finally, special thanks to my boyfriend Greg and my cat Bobo for always being on my side.

## Table of Contents

Abstract.....	ii
1 Introduction .....	1
2 Approach.....	2
2.1 Sample characteristics of the TS and WGS data.....	2
2.2 Data generation and variant calling.....	3
2.3 Data cleaning .....	3
2.4 Comparative analysis.....	4
2.4.1 Comparison of variants sets .....	4
2.4.2 Concordance of genotype calls .....	4
2.4.3 Genotypic TDT .....	5
2.4.4 Linkage disequilibrium (LD) .....	6
3 Results.....	7
3.1 Data cleaning .....	7
3.2 Comparative Analysis.....	7
3.2.1 Comparison of Variants Sets .....	7
3.2.2 Concordance of genotype calls .....	10
3.2.3 Genotypic TDT and linkage disequilibrium comparison .....	13
4 Summary .....	16
5 Bibliography .....	20
6 Appendix.....	21
7 CV .....	22



## List of Tables

<b>Table 1.</b> Sample size in number of individuals and trios for the TS and WGS data. (The TS data had already been cleaned and processed as part of another study, so no further individual-level cleaning was done here. Variant-level cleaning was performed similarly across the two data sets.).....	3
<b>Table 2.</b> Breakdown for genotype call mismatches between the WGS data and the TS data in 134 trios and 7,183 variants by physical position on 8q24 (hg19).....	10
<b>Table 3.</b> Breakdown for genotype call mismatches between the WGS data and the TS data in 134 duplicated trios for 7183 variants by individuals on 8q24 (hg19).....	11
<b>Table 4.</b> Summary statistics for average read depth (DP) based on mismatch type in genotype calls (GT) for the WGS data in 402 individuals and 7183 variants.....	12
<b>Table 5.</b> Summary statistics for average call quality (GQ) based on mismatch type in genotype calls (GT) for the WGS data in 134 trios and 7,183 variants. ....	13
<b>Table 6.</b> Number of common variants based on each filtering step prior to SNP plot generation for the TS, WGS and the combined data .....	13

## List of Figures

<b>Figure 1.</b> Histogram of Mendelian Error counts by family in 332 trios from the WGS study .....	7
<b>Figure 2.</b> Venn Diagram for number of distinct SNVs in the parents. Common variants in 654 individuals from the WGS study (green) vs. common variants in 480 individuals from the TS study (pink). .....	8
<b>Figure 3.</b> Venn Diagram for number of distinct SNVs in the parents. Rare variants ( $MAF < 0.01$ ) in 480 individuals from the TS study (pink) vs. all variants in 654 individuals from the WGS study (green). Of the 18,854 rare variants seen only in the TS study, 2,967 (15.7%) were singletons.....	8
<b>Figure 4.</b> Venn Diagram for parents-only data. Common variants in 480 individuals from the TS study (pink) vs. common variants in 654 individuals from the WGS study (green). .....	9
<b>Figure 5.</b> Venn Diagram for number of distinct SNVs in the parents. Common variants in 480 individuals from the TS study (pink) vs. rare variants ( $MAF < 0.01$ ) in 654 individuals from the WGS study (green). Of the 10,244 rare variants seen only in the WGS study, 3,881 (37.9%) were singletons.....	9
<b>Figure 6.</b> Histogram for genotype call mismatches between the WGS data and the TS data in 134 trios and 7,183 variants by physical position on 8q24 (hg19). .....	10
<b>Figure 7.</b> Histogram for genotype call mismatches between the WGS data and the TS data in 134 duplicated trios for 7,183 variants on 8q24 (hg19) by individuals. ....	11
<b>Figure 8.</b> Histogram for average read depth (DP) by individual based on mismatch type in genotype calls (GT) for the WGS data in 402 individuals and 7183 variants. ....	12
<b>Figure 9.</b> Histogram for average call quality (GQ) based on mismatch type in genotype calls (GT) for the WGS data in 134 trios and 7,183 variants. ....	13
<b>Figure 10.</b> SNP plot for common variants ( $MAF \geq 0.01$ ) based on WGS study (327 trios, 276 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 8. ....	14
<b>Figure 11.</b> SNP plot for common variants ( $MAF \geq 0.01$ ) based on TS Study without duplicates (240 trios, 215 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 7. ...	14
<b>Figure 12.</b> SNP plot for common variants ( $MAF \geq 0.01$ ) based on the combined data set of TS Study and WGS study (567 trios, 210 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 14. ....	15
<b>Figure 13.</b> gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the WGS Study without duplicates (327 trios, 3161 common variants) for all positions on 8q24. Note that y-axis range is from 0 to 8. ....	21
<b>Figure 14.</b> gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the TS Study without duplicates (240 trios, 2559 common variants) for all positions on 8q24. Note that y-axis range is from 0 to 7. ....	21
<b>Figure 15.</b> gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the combined data without duplicates (567 trios, 2481 common variants) for all positions on 8q24. Note y-axis range is from 0 to 14. ....	21

# 1 Introduction

Congenital anomalies are one of the most common causes of infant and childhood mortality worldwide; among these birth defects, orofacial clefts (OFCs), including cleft lip (CL), cleft palate (CP) as well as cleft lip and palate (CLP), are the most common group of craniofacial malformations with a birth prevalence rate of around 0.17% (TOLAROVA 2018). Children born with orofacial clefts also have an increased risk of developing mental problems and cancers (LESLIE AND MARAZITA 2013). Considering the public health burden caused by OFCs, study and research into this field should be valuable in improving population health.

Orofacial clefts can be further classified into *syndromic* oral clefts which occur with other etiologically or pathogenically related malformations and *non-syndromic* OFCs (MOSSEY AND CASTILLA 2001). The primary focus of this study and previous studies by our group has been on isolated, *non-syndromic* OFCs to examine potential genetic variants that may be associated with this common birth defect. Based on previous GWAS (genome-wide association study) results, the 8q24 region, ranging from 129,778,467 to 130,181,350 on chromosome 8, has yielded significant evidence of association with CL/P (BIRNBAUM *et al.* 2009; GRANT *et al.* 2009), and this statistical significance is stronger in European populations compared to Asians (BEATY *et al.* 2010). An additional study using case-parent trio data from 13 selected genetic regions has further confirmed these population-specific genetic patterns (LESLIE *et al.* 2015), which may in part be explained by differences in SNP heterozygosity between European and Asian trios (MURRAY *et al.* 2012).

Despite the strong statistical evidence for association between genetic variation in the 8q24 region and development of an OFC, this region is a gene desert and contains very few recognized genes, so predicting functional variants is difficult (HUPPI *et al.* 2012). Cross-population differences in association signal provide one avenue for narrowing down the potential functional region; one hypothesis is that truly causal variants should maintain functional roles across populations even when association signal differs due to differences in correlation or linkage disequilibrium (LD) patterns across different populations. We have recently collected new whole-genome sequencing (WGS) data on a set of 1,136 European ancestry individuals (in approximately 378 case-parent trios) as part of the Gabriella Miller Kids First Pediatric Research Program, a trans-NIH effort currently focused on gene discovery for pediatric cancers and structural birth defects (<https://commonfund.nih.gov/KidsFirst>). In addition, we have existing targeted sequencing (TS) data of this 8q24 region on 374 European case-parent trios (of which 134 are duplicated in the WGS data) and 1,034 Asian trios, which were described analyzed previously (LESLIE *et al.* 2015).

The goal of our current analysis is to build on this prior work by leveraging the increased sample size provided by the new WGS data in the hope of further refining the association signal in this region, and thereby potentially illuminating functional mechanisms that may lead to the development of OFCs. However, prior to combining these data sets, which were generated at different time points using different technologies, we first need to test for systematic differences in quality and variant calls between the TS and WGS data sets. This comparison will be performed by assessing coverage of rare and common variants in the independent case-parent trios across the TS and WGS data sets, and examining concordant and mismatched genotype calls for a sub-set of 402 individuals who were sequenced in both the TS and WGS studies, in terms of their call quality and read depth. Once this assessment has been performed, we will compare evidence of linkage and association obtained from these case-parent trios, comparing results from the TS data, the WGS data and a combined data set incorporating both sample sets. Specifically, the statistical signals from the genotypic transmission-disequilibrium test (gTDT) and measures of haplotype diversity will be compared between the TS and WGS data to check for consistency of called variants in this 8q24 region. This combined analysis provides a resource for further understanding the functional genetics of the 8q24 region, with future work planned to assess cross-population differences with this larger harmonized data set. The ultimate purpose of this study was to combine targeted and WGS data to gain more statistical power and accuracy for future studies, not only in the 8q24 region, but throughout the genome.

## **2 Approach**

### **2.1 Sample characteristics of the TS and WGS data**

Samples for the TS data set were collected from individuals of Asian or European ancestry from Europe, the United States, China, and the Philippines recruited due to the presence of a child affected with cleft lip (CL) or cleft lip with cleft palate (CLP), collectively referred to as CL/P. Individuals with other congenital anomalies, recognized malformation syndromes involving CL/P, or developmental delays were excluded from the study. For our work here, we are focusing only on those families of European ancestry. Samples for the WGS data set were drawn from individuals who were part of the same studies, but who are exclusively of European origin. A total of 402 individuals (in 134 case-parent trios) from the WGS study were also included in the TS study. These overlapping trios create a unique opportunity for direct comparison of variants identified by both sequencing methods, so we compared the called variants in this sub-set for call rates and consistency. However, these overlapping samples were only retained in the WGS data for our analyses of linkage and association.

**Table 1.** Sample size in number of individuals and trios for the TS and WGS data. (The TS data had already been cleaned and processed as part of another study, so no further individual-level cleaning was done here. Variant-level cleaning was performed similarly across the two data sets.)

Data	# of individuals (trios) before data cleaning	# of individuals (trios) after data cleaning
TS Data	NA	1122 (374)
TS Data (w/o duplicates)	NA	402 (134)
WGS Data	1136 (~379)	981(327)

## 2.2 Data generation and variant calling

Sample preparation and sequencing for both the TS and WGS data sets was performed at the McDonnell Genome Institute (MGI) of Washington University. For the TS study, thirteen high priority regions were selected for sequencing, representing 6.3Mb. In brief, NimbleGen (Roche NimbleGen, Madison, WI) custom target probes were designed to query the 6.6Mb target region, and hybrid capture on pools of 96 indexed samples per capture was performed. Each capture pool was then sequenced on two lanes of Illumina HiSeq per manufacturer's recommendations (Illumina Inc, San Diego, CA) for an average of ~40Gb per lane or ~835Mb per sample. For the WGS study, data generated were 2x150bp paired end reads with a target of ~30X coverage. In both cases, MGI applied state of the art alignment and variant calling routines, including variant/sample QC (flagged for per-sample call rate, per-marker call rate and adherence to Hardy-Weinberg equilibrium among parents) and pedigree-aware genotype refinement using Polymutt (Li *et al.* 2012). Here, we focus on the 8q24 region (hg19 genome coordinates 129,778,467 to 130,181,350) which was previously shown through GWAS to be associated with risk of CL/P, and which was replicated in the original TS main publication (LESLIE *et al.* 2015).

## 2.3 Data cleaning

Prior to the comparative analysis, the WGS data were cleaned to maximize consistency for this study. First, individuals from incomplete trios (i.e., those missing one or both parents) were excluded and only complete trios were used for our analysis. Second, multi-allelic SNPs were removed and genotype calls were filtered based on read depth and call quality: calls with read depth (DP) less than 10 or call quality (GQ) less than 20 were set to missing. Then,

incomplete calls with a missing value for one allele and a non-missing value for the other were set to missing as well (these “half-calls”, e.g., ./1 in the VCF file). Since this study is focused on the association signal from the 8q24 region, this specific region was subset from chromosome 8 and tests for Mendelian consistency were conducted to check for inheritance errors within each trio. Additionally, the results from this Mendelian error check in chromosome 8 were compared with the results for all variants identified on chromosome 22 to confirm their validity. Families that were outliers in their count of Mendelian errors compared to the distribution of all other families were deleted.

## **2.4 Comparative analysis**

### **2.4.1 Comparison of variants sets**

To examine the overlap of variants called in the TS and WGS data sets, variants were divided into classes according to their within-data set minor allele frequency (MAF), calculated in parents only: singletons (variants only seen in one individual in the data set), rare variants ( $MAF < 0.01$ ) or common variants ( $MAF \geq 0.01$ ). Venn Diagrams were generated to directly visualize the overlap among these classes of variants between the TS and WGS data sets. As the goal was to examine the overlap rate as a measure of technical reproducibility, the 402 duplicated individuals were removed from the TS data. These Venn Diagrams were graphed separately for common and rare variants and for each data set, e.g., rare variants from the WGS data were plotted with all variants from TS data set to see how many of these rare variants were also present in the TS data set at any frequency. For non-overlapping rare variants, the total number of singleton variants was calculated to assess the impact of singletons on whether a variant failed to appear in both data sets, since singleton variants are perhaps the least likely to overlap between data sets.

### **2.4.2 Concordance of genotype calls**

For the 402 individuals duplicated between the TS and WGS data sets, we examined the consistency of genotype calls (GT) for all overlapping positions. Only variants at positions present in both data sets were kept. Then, counts of mismatches in genotype calls were calculated and plotted separately by individual and by position; in other words, mismatches were obtained based on both rows (position) and columns (individual). To examine if patterns of read depth (DP) and call quality (GQ) were associated with mismatches in genotype calls, the average read depth and call quality were calculated for each individual based on the mismatch tag (either this call has a mismatch between these two data sets or not) and the patterns were visualized using histograms.

### 2.4.3 Genotypic TDT

The Transmission-Disequilibrium Test (TDT) aims to test the composite null hypothesis of no linkage or no LD between an observed marker and an unobserved causal locus to detect over-transmission of a particular marker allele from the expected probability of  $\frac{1}{2}$  that any given allele is transmitted from a heterozygous parent to an affected offspring at meiosis (SPIELMAN AND EWENS 1996). In contrast to the allelic TDT, which uses McNemar's chi-squared test to assess the transmission of alleles, the genotypic TDT utilizes conditional logistic regression to examine the transmission of genotype to the affected child (LAIRD AND LANGE 2006). This allows for more flexible modeling of the genetic effect at the unobserved causal locus (i.e., additive, dominant or recessive models of inheritance) and produces estimated odds ratios and confidence intervals for each genotype as a default. Specifically, the genotypic TDT considers the observed genotype in the case from each trio, plus three unobserved "pseudo-controls" which represent all possible genotypes of children from the parental mating, as a cluster, and models a case-control test of association within each cluster in a conditional logistic regression framework. Since there are four possible combinations of genotypes for the child based on parents' mating type, the genotype of the affected offspring becomes a case while the other three possible genotypes are treated as pseudo-controls. The null hypothesis in this setting states that there is no association between observed genotype and disease and the marker is not linked to the causal locus. If the probability for the observed genotype is significantly larger than expected due to rules of Mendelian inheritance, the composite null hypothesis is rejected, and there is evidence of linkage and association between the observed marker and some unobserved gene controlling the disease of interest (LAIRD AND LANGE 2006). In our analysis, only common variants with MAF larger than or equal to 1% are considered for the genotypic TDT, as the statistical power of drawing inference using rare variants is limited.

Despite the QC steps mentioned in the data cleaning section, additional filtering steps were conducted for the calculation of the gTDT and further plotting. First, the duplicated individuals (in 134 case-parent trios) were removed from the TS data set, so we were comparing signals among two mutually-exclusive groups. Then, the TS and WGS data sets, which were in vcf format, were converted into matrices in genotype format by combining the information from the pedigree files; note that variants with MAF of less than 1% were filtered out so we only focused on analyzing signals from common variants. Additionally, SNPs with more than 5% missing calls and those showing significant deviation from Hardy-Weinberg equilibrium (at p-value less than 0.05) were filtered out as well. Finally, the gTDT was calculated based on the cleaned genotype matrix and missing values in gTDT results were also removed. These

gTDT plots were first generated based on all markers in the 8q24 region to observe the overall patterns among all positions, and then a subset of regions containing the strongest signals were selected for further analysis.

To further examine the consistency and reproducibility in signals of linkage and association, the TS and WGS data sets were merged based to all variants seen in the two data sets. This combined data set was processed following the same filtering procedures as mentioned above, and a gTDT plot was generated based on all positions in 8q24 first to visualize the overall pattern in signals of linkage and association, and then subset to the region with the strongest signals.

#### 2.4.4 Linkage disequilibrium (LD)

Linkage disequilibrium (LD) is a measure of correlation between alleles at different genetic markers or SNPs. It is typically higher for variants that are physically close to one another on a chromosome. Patterns of LD can vary between genomic regions and between populations. LD is a measure of whether alleles at different SNPs can be combined based on the Hardy-Weinberg expectations within a population. Specifically, assume allele  $A_1$  and allele  $A_2$  at a SNP A occur at frequencies of  $p_1$  and  $p_2$ , respectively, and  $p_1 + p_2 = 1$ . When mating is random with respect to genotype, Hardy-Weinberg expectations predict frequencies of  $A_1A_2$ ,  $A_1A_1$  and  $A_2A_2$  are  $2p_1p_2$ ,  $p_1^2$  and  $p_2^2$ , respectively. Similarly, assume allele  $B_1$  and allele  $B_2$  at a SNP B occur at frequencies of  $q_1$  and  $q_2$ , respectively, where  $q_1 + q_2 = 1$ , with random mating. The two SNPs are said to be in linkage equilibrium if the frequencies for the four possible gametes  $A_1B_1$ ,  $A_1B_2$ ,  $A_2B_1$ ,  $A_2B_2$  are  $p_1q_1$ ,  $p_1q_2$ ,  $p_2q_1$  and  $p_2q_2$ , respectively. On the contrary, the two SNPs are said to be in linkage disequilibrium if these same combinations of alleles at two loci (i.e. these four haplotypes) cannot be predicted from their respective allele frequencies (HARTL AND CLARK 1997).

Such LD can be measured by the correlation coefficient between alleles at two different SNPs which is calculated as follows (ULEBERG AND MEUWISSEN 2011):

1. Let  $p_1$  denote the frequency of allele  $A_1$  at SNP A and  $q_1$  denote the frequency of allele  $B_1$  at SNP B.
2. Let  $p$  denote the observed frequency of the  $A_1B_1$  haplotype.
3. The deviation between the expected haplotype frequency is therefore defined as:  $D_{A_1B_1} = p - p_1q_1$
4. Calculate the correlation coefficient as:  $r = \frac{D_{A_1B_1}}{\sqrt{p_1(1-p_1)q_1(1-q_1)}}$
5. Obtain the coefficient of determination by taking the square of the correlation coefficient:  $r^2$

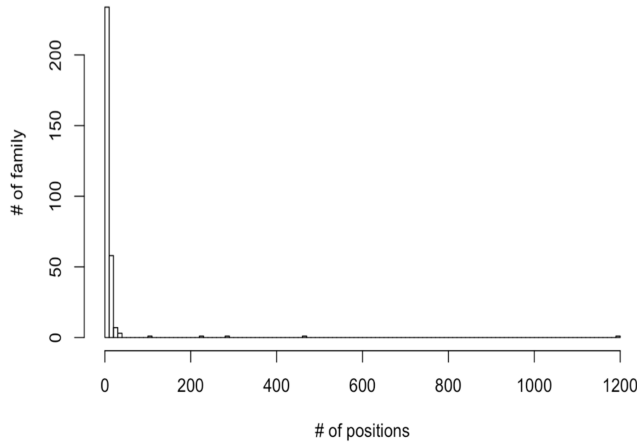
As mentioned in **Section 2.4.3**, we calculated LD ( $r^2$ ) for variants in the filtered data sets from the TS, WGS and the combined data. Heat maps were generated based on these  $r^2$  values and the results were mapped with their positions



in hg19.

## 3 Results

### 3.1 Data cleaning



**Figure 1.** Histogram of Mendelian Error counts by family in 332 trios from the WGS study

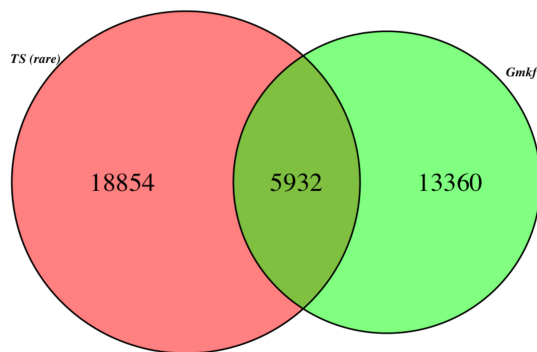
variants; but there were 5 families with significantly higher numbers of Mendelian errors compared to the others (range 104-1,194 variants), and these 5 families were removed from the WGS study.

As presented in **Table 1**, there were 1,136 individuals in the original WGS data and 981 individuals (in 327 complete trios) after the data cleaning described above. In the TS data, there were 1,122 individuals (in 374 complete trios of European ancestry). Of these, 134 case-parent trios were also present in the WGS data set, leaving 240 trios for the analysis of independent subjects between these two data sets.

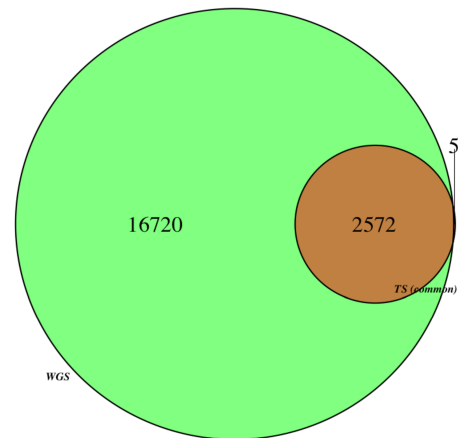
### 3.2 Comparative Analysis

#### 3.2.1 Comparison of Variants Sets

After the QC steps mentioned as above, 332 trios for the WGS data remained for analysis. Tests for Mendelian consistency were conducted on each family. **Figure 1** presents a histogram of Mendelian errors based on all families. The x-axis indicates the number of positions, while, the y-axis indicates the number of families. Based on **Figure 1**, the vast majority of families had Mendelian errors at fewer than 50

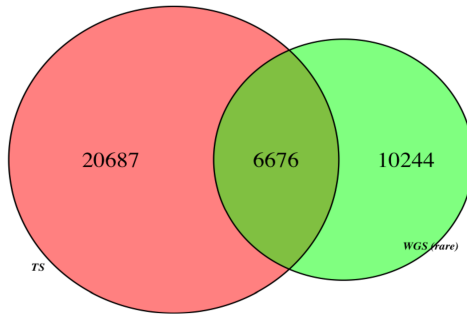


**Figure 3.** Venn Diagram for number of distinct SNVs in the parents. Rare variants (MAF < 0.01) in 480 individuals from the TS study (pink) vs. all variants in 654 individuals from the WGS study (green). Of the 18,854 rare variants seen only in the TS study, 2,967 (15.7%) were singletons.

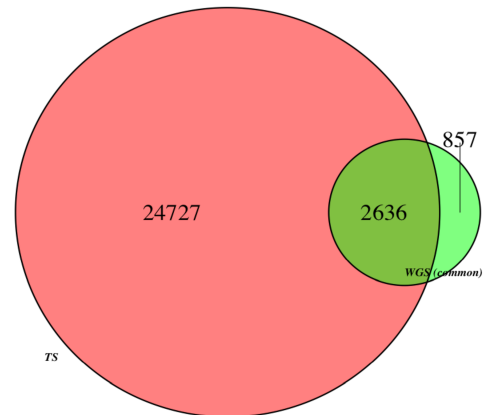


**Figure 2.** Venn Diagram for number of distinct SNVs in the parents. Common variants in 654 individuals from the WGS study (green) vs. common variants in 480 individuals from the TS study (pink).

**Figures 2 and 3** present the Venn Diagrams for all variants in the WGS data compared to the TS data for rare and common variants, respectively. Note that both of these figures represent only parents' information and are mutually exclusive; the WGS data had 480 individuals, while the targeted data set had 654 individuals. As illustrated in **Figure 2**, only 5,932 rare variants (MAF < 0.01) out of a total number of 24,786 in the TS data (23.9 %) were also found in the WGS data. Out of the 18,854 non-overlapping rare variants in the TS data set, 2,967 (15.7 %) were singletons, i.e. found only in one individual. In **Figure 3**, the majority of common SNPs in the TS data were also found in the WGS data; and only 5 common SNPs out of a total number of 2,577 in TS data (0.02 %) were seen only in the TS data.



**Figure 5.** Venn Diagram for number of distinct SNVs in the parents. Common variants in 480 individuals from the TS study (pink) vs. rare variants (MAF < 0.01) in 654 individuals from the WGS study (green). Of the 10,244 rare variants seen only in the WGS study, 3,881 (37.9%) were singletons.



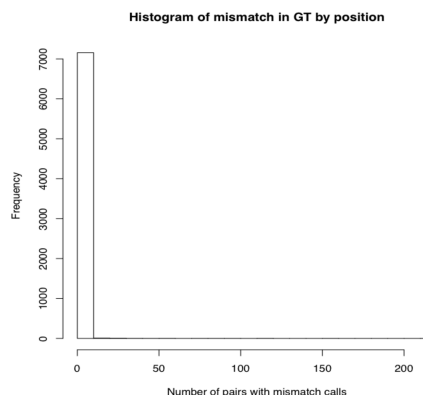
**Figure 4.** Venn Diagram for parents-only data. Common variants in 480 individuals from the TS study (pink) vs. common variants in 654 individuals from the WGS study (green).

**Figures 4 and 5** represents similar Venn Diagrams for all variants in the TS data compared to the WGS data for both rare and common variants, respectively. Similar patterns were observed in **Figure 4**: only 6,676 rare variants out of a total number of 16,920 (39.5%) in the WGS data were also found in the TS data. Among the 10,244 non-overlapping rare variants seen in the WGS data, 3,881 (37.9%) were singletons. Similarly, **Figure 5** shows the majority of common SNPs in the WGS data were also found in TS data but the overlapping percentage was smaller compared to the pattern seen in **Figure 3**; 857 common variants out of a total number of 3,493 (24.5%) in the WGS data were also in the non-overlapping region whereas only 5 common variants out of 2,577 in the TS data (0.02%) did not overlap. This pattern indicates many common variants were captured by the WGS data were missed by the TS approach.

### 3.2.2 Concordance of genotype calls

**Table 2.** Breakdown for genotype call mismatches between the WGS data and the TS data in 134 trios and 7,183 variants by physical position on 8q24 (hg19).

Count (%) of pairs with mismatch calls	Count of positions
0 (0%)	6564
1-10 (0.25%-2.5%)	592
11-50 (2.7%-12.4%)	18
51-100 (12.7%-24.9%)	5
101-150 (25.1%-37.3%)	2
151-200 (37.6%-49.8%)	1
201-250 (50%-62.19%)	1
251+ (62.4%+)	0



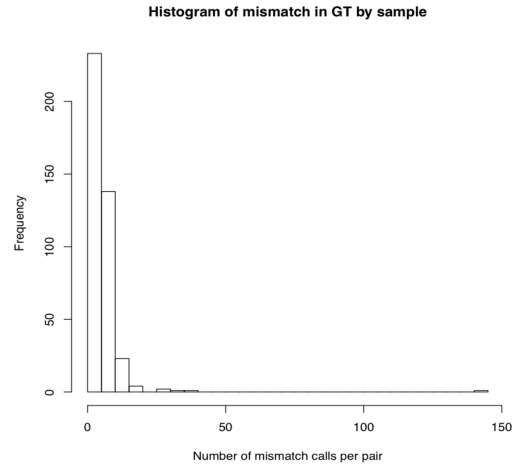
**Figure 6.** Histogram for genotype call mismatches between the WGS data and the TS data in 134 trios and 7,183 variants by physical position on 8q24 (hg19).

To assess how well genotype calls matched between the TS and WGS data sets, we carefully examined the set of 402 individuals who were sequenced in both studies. We focused on variants present in both data sets regardless of MAF, of which there were 7,183 SNVs, and checked for concordance between individual genotype calls at these positions in the duplicated pairs. Each genotype call was classified as a “match” or a “mismatch” depending on whether the same genotype call was made in both data sets, or not. **Figure 6** presents the histogram of counts of mismatched genotype calls by position. In other words, for each position, the number of mismatches in genotype calls were obtained across all 402 pairs of duplicated individuals. The x-axis shows the number of pairs with mismatched variant calls, while the y-axis indicates their frequency. Based on the distribution seen in **Figure 6**, clearly the majority of positions had fewer than 10 mismatched calls (< 2.5% in terms of error rate).

**Table 2** further presents details for this breakdown. Out of a total of 7,183 observed SNPs, there were 7,156 positions showing a mismatch number of 10 or less, which is 2.5% in terms of error rate across all duplicate pairs of individuals. Only 9 positions had a mismatch rate of 12.7% or larger. In general, the majority of variant calls in the two data sets matched for each position, and most variants only had a small number of mismatched pairs.

**Table 3.** Breakdown for genotype call mismatches between the WGS data and the TS data in 134 duplicated trios for 7183 variants by individuals on **8q24 (hg19)**.

Count (%) of mismatch calls per pair	Count of pairs
0 (0%)	2
1-5 (0.01%-0.06%)	230
6-10 (0.08%-0.14%)	138
11-20 (0.15%-0.28%)	27
21-30 (0.3%-0.42%)	2
31-40 (0.43%-0.56%)	2
41-140 (0.57%-1.94%)	0
141 (1.96%)	1
142+ (1.98%+)	0



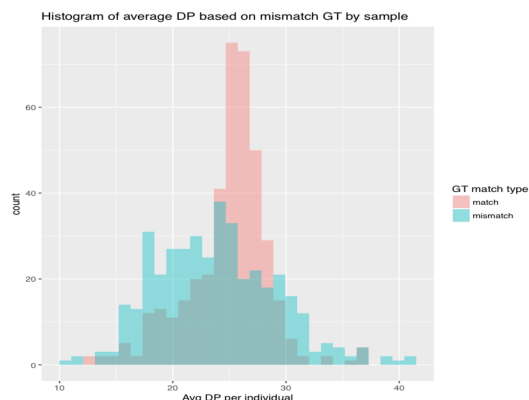
**Figure 7.** Histogram for genotype call mismatches between the WGS data and the TS data in 134 duplicated trios for 7,183 variants on 8q24 (hg19) by individuals.

**Figure 7** displays the histogram of mismatches in genotype calls for each duplicate pair of individuals. In other words, for each individual, the number of mismatched genotype calls was counted across all 7,183 positions. Again, the x-axis indicates the number of mismatched calls per pair, while the y-axis indicates its frequency. Based on the distribution seen in **Figure 7**, the vast majority of the pairs had a mismatch number of less than 30.

**Table 3** presents additional details for the counts of mismatched genotype calls by each duplicate pair of individuals. Out of a total number of 402 duplicate pairs of individuals, there were 399 pairs having 30 or fewer mismatched calls, which is 0.42% in terms of error rate across all 7,183 positions regardless of MAF. Additionally, only 1 duplicated pair had a mismatch rate larger than 1%. In general, the majority of genotype calls in these two data sets matched very well in these duplicated pairs of individuals.

**Table 4.** Summary statistics for average read depth (DP) based on mismatch type in genotype calls (GT) for the WGS data in 402 individuals and 7183 variants.

Summary Statistics for Average DP	Match in GT	Mismatch in GT
Median	25.4	23.8
Mean	24.9	23.9
Standard Deviation	3.6	5.3

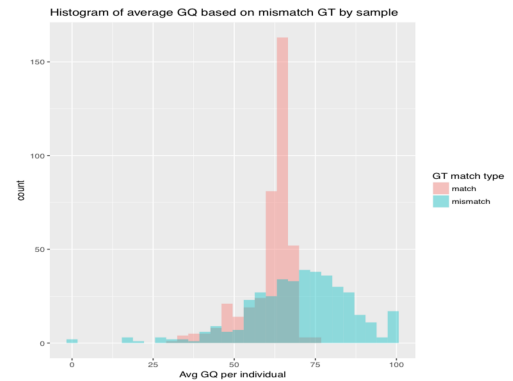


**Figure 8.** Histogram for average read depth (DP) by individual based on mismatch type in genotype calls (GT) for the WGS data in 402 individuals and 7183 variants.

**Figure 8** shows the histogram for average read depth (DP) within the WGS data for two classes of variants within each individual: those with calls matching between the WGS and TS data sets and those where the called genotypes did not match. DP is a measure of coverage that quantifies the number of unique sequencing reads containing a specified nucleotide. Within each class of variants, the average read depth for each individual was calculated and then plotted. The blue color indicates the average for positions with a mismatch in genotype calls, while the red color indicates the average for positions where the genotype call matched. In general, the distribution of average read depth for the mismatched group has a larger standard deviation and a smaller mean compared to the distribution for the matched group, as expected. The summary statistics from **Table 4** further validate the patterns being observed in the histogram in **Figure 8**. Based on **Table 4**, the standard deviation for the mismatched group was larger compared to the matched group (5.3 vs. 3.6), while the mean for the mismatched group was smaller (23.9 vs. 24.9). These results make technical sense: Positions with lower depth are more prone to erroneous calls as there is less information from which to call the genotypes. The mismatch calls are also likely a mix between correct and incorrect calls, which might contribute to the higher spread in DP measures (typical of a mixture distribution).

**Table 5.** Summary statistics for average call quality (GQ) based on mismatch type in genotype calls (GT) for the WGS data in 134 trios and 7,183 variants.

Summary Statistics for Average GQ	Match in GT	Mismatch in GT
Median	63.6	71.6
Mean	61.0	70.1
Standard Deviation	7.4	16.0



**Figure 9.** Histogram for average call quality (GQ) based on mismatch type in genotype calls (GT) for the WGS data in 134 trios and 7,183 variants.

**Figure 9** is the histogram of the average call quality (GQ) based on the genotype calls that matched or did not match for the 402 individuals in the WGS data. Within each class of variants, the call qualities for each sample were averaged and then plotted. GQ is a measure provided by the genotype calling software and provides a numeric value representing how certain the caller was about the call being made. The distribution of average call quality for the mismatched group had a larger standard deviation, but a higher mean compared to the distribution for the matched group, patterns quite different from that seen in the read depth. The summary statistics from **Table 5** further show the standard deviation for the mismatched group was much larger than in the matched group (16 vs. 7.4) and the mean for the mismatched group was larger as well (70.1 vs. 61). The reasons for this higher mean in the mismatched group are not clear; the higher standard deviation can again be attributed to the fact that these variants are likely a mix of correct and incorrect calls.

### 3.2.3 Genotypic TDT and linkage disequilibrium comparison

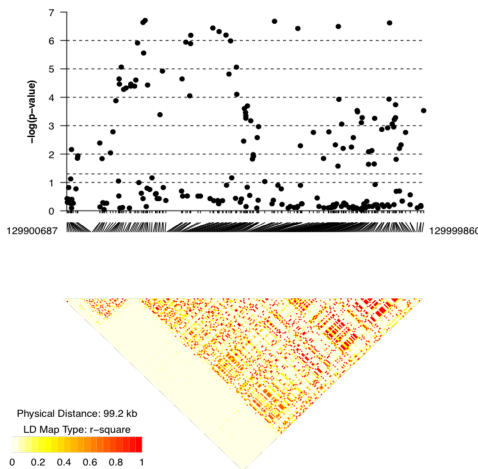
**Table 6.** Number of common variants based on each filtering step prior to SNP plot generation for the TS, WGS and the combined data

Variant Filtering Steps	TS data	WGS data	Combined Data
Total # of common variants in the Geno file	3209	3393	2731
# of SNPs with over 5% missing calls	562	124	122
# of SNPs with p values < 0.05 for HWE	81	97	123
# of SNPs with NAs in gTDT	7	11	5
Total # of variants after all filtering steps for 8q24 regions	2559	3161	2481
Total # of variants after all filtering steps for	215	276	210

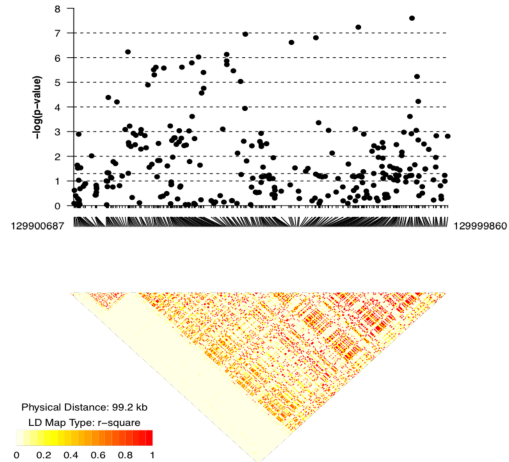
the sub-set regions (129,900,000 -  
130,000,000)

As mentioned in the Approach section, some filtering steps were conducted prior to the plotting of the results of the genotypic TDT and the LD heat map. **Table 6** specifies the number of common variants removed at each filtering step. In the original genotype matrix obtained from the TS data and its pedigree file, there were 3,209 common variants; after filtering based on missing calls, p-values from Hardy-Weinberg Equilibrium and missing values in genotypic TDT, there were 2,559 common variants left for the full 8q24 region. Similarly, for the WGS data, there were 3,393 common variants in the initial genotype matrix; after filtering, we were left with 3,161 common variants for plotting. In terms of the combined data set, the genotype matrices of the TS and WGS data were merged based on the overlapping variants and then this merged matrix was filtered based on the same pipeline. There were 2,731 common variants for the initial genotype matrix of the combined data set and after filtering, we were left with 2481 common variants for the 8q24 regions.

**Figures 13, 14 and 15** in the appendix illustrate the gTDT results based on all 8q24 variants for the TS (240 trios), WGS (327 trios) and the combined (567 trios) data sets, respectively. Both plots from the TS and the WGS studies were generated from trios of European ancestry and those trios were mutually exclusive. These plots represent  $-\log_{10}(p)$  for the genotypic Transmission Disequilibrium Test (gTDT) where  $-\log_{10}(p)$  was plotted along all positions in 8q24 region (in hg19). In general, the signal of over-transmission was much stronger in the combined data



**Figure 11.** SNP plot for common variants ( $MAF \geq 0.01$ ) based on TS Study without duplicates (240 trios, 215 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 7.

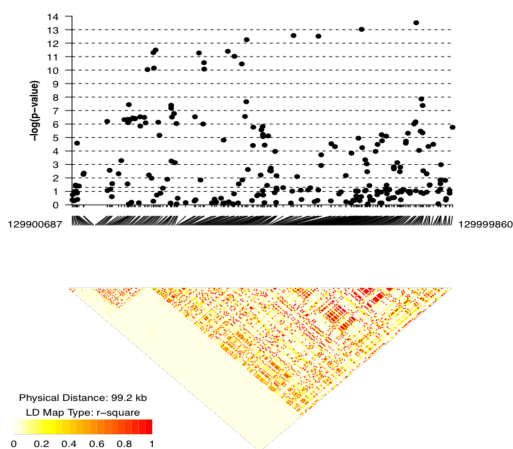


**Figure 10.** SNP plot for common variants ( $MAF \geq 0.01$ ) based on WGS study (327 trios, 276 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 8.



set, as the maximum value of y-axis is larger. Additionally, more common variants ( $MAF \geq 0.01$ ) were detected in the WGS data for the entire 8q24 region (3,161 common variants).

To further examine the sub-region of 8q24 with the strongest signal in the gTDT, we selected all positions falling between 129,900,000 and 130,000,000 on hg19, based on the location of the strongest signal in **Figures 13, 14**



**Figure 12.** SNP plot for common variants ( $MAF \geq 0.01$ ) based on the combined data set of TS Study and WGS study (567 trios, 210 common variants) for the sub-set region in 8q24 with hg19 genome coordinates ranged from 129,900,000 to 130,000,000. Note y-axis range is from 0 to 14.

**Figure 12** illustrates the SNP plot for the combined data set by merging the TS data with the WGS data based on the overlapping regions. The combined data set contains 567 trios and 210 common variants. Based on **Figure 12**, the signal of over-transmission is much stronger compared to **Figures 10** and **11** likely due to larger sample size and a concordance of genetic background due to the common European ancestry of all samples included in this analysis.

In **Figures 10-12**, the bottom panel represents a heat map based on pairwise LD as measured by  $r^2$ , which is a measure of the correlation between alleles at different pairs of SNPs in a population. An  $r^2$  value of zero indicates that the allele at one position does not have any association with the other allele typically due to genetic recombination between the positions. For all the figures, the LD patterns were quite similar; there were two unequal-sized triangular blocks with the boundary position at a genetic recombination site. Even for the combined data set, little additional haplotype structure was observed at this scale, although there is a small “sub-block” within the large block that shows increased  $r^2$  values. This slight increase in detail may help with further refining the gTDT signal in this region to better

and **15**. As shown in **Table 7**, there were 215, 276 and 210 common variants left for the sub-region in the TS, WGS and the combined data, respectively. **Figure 10** presents the SNP plot based on the sub-region for the TS data (240 case-parent trios), while **Figure 11** shows a similar SNP plot for the WGS data (327 case-parent trios). Similar to the patterns observed in **Figures 13** and **14**, the signal of over-transmission is slightly stronger in the WGS data, so the maximum value of the y-axis is larger. Additionally, more common variants ( $MAF \geq 0.01$ ) were detected in WGS data compared to the TS data (276 common variants vs. 215 common

identify potentially causal variants.

## 4 Summary

Briefly speaking, the purpose of this study was to compare the TS data obtained from a previous study and the WGS data generated recently and to propose a reasonable approach to combine these two data sets thereby gaining more statistical power and accuracy for future analysis. In terms of detailed approaches, Venn Diagrams were generated based on common ( $MAF \geq 0.01$ ) and rare variants ( $MAF < 0.01$ ) in each data set to further examine and visualize the overlap rate of variants called in the two data sets; the number of singletons (variants only seen in one individual in the data set) was calculated for the non-overlapping rare variants in each data set to monitor the impact of singletons on the overlap rate for rare variants. The consistency of genotype calls was examined based on the 402 duplicated individuals appearing in both data sets and overlapping positions regardless of MAF; in addition, histograms were generated for average read depth (DP) and call quality (GQ) based on the mismatched tag (whether the variant calls were different for the same individual and same position among the two data sets or not) to monitor any different patterns occurring due to mismatched calls. Lastly, genotypic TDT plots and LD heat maps were generated for the TS, WGS and the combined data sets to examine if the signals of linkage and association were consistent and reproducible.

After the general data cleaning step, we had 981 individuals (in 327 case-parents trios) left for the WGS study and 1,122 individuals (in 374 case-parents trios) left for the TS study; among these 1,122 individuals for the TS study, 402 individuals (in 134 case-parents trios) were present in both data sets. For the analysis of comparing variants sets, only parents' data were kept for calculating minor allele frequency and all duplicated individuals were removed from the TS study. Thus, we had 654 individuals left for the WGS study and 480 individuals left for the TS study. Based on the results from the Venn Diagrams, almost all common variants (99.8%) from the TS study were present in the WGS study and the majority of common variants (75.5%) from the WGS study were also present in the TS study. In other words, the techniques used by the WGS study captured more common variants compared to that of the TS study. This may be due, for example, to capture inefficiency in the TS data generation process, or differences in coverage due to batch effects. Overall, combining these two data sets based on overlap in common variants is a reasonable approach for further study as the patterns are largely concordant.

In terms of the overlap rate for rare variants, only 23.9% of the rare variants from TS study were present in the WGS study, and 39.5% of the rare variants from the WGS study were also found in the TS study. The reasons for

this are potentially more complex. Since a MAF of 1% corresponds to fewer than 7 individuals in the WGS data or 5 individuals in the TS data, these differences may be due to random sampling. They may also be due to technical challenges with calling rare variants. To explain these observed differences, we examined the number of non-overlapping variants that were singletons, i.e., that only appeared in one individual. Generally, rare variants were singletons in one data set may not be present at all in the other data set because the variant was too rare to be detected in another sample. Moreover, for each additional person in a data set, the number of singletons will likely increase, which may explain the higher rate of singletons in the larger WGS data set: A significant portion (37.9%) of the non-overlapping rare variants in the WGS data set were singletons; however, the impact of singletons in the TS data set (15.7% of non-overlapping variants) was not as strong. For rare variants, it is quite possible that some of the rare variants and possibly a large fraction of the singleton variants actually represents artifacts, so it might be expected that direct merging of the two data sets based on rare variants may not be an informative approach. On the other hand, since our goal is to eventually determine functional, deleterious mutations in this region that could contribute to disease etiology, and since rare variants may be particularly likely candidates for this role, increasing the number of rare variants in our data set may greatly increase our power to detect this functional signal.

The findings from **Section 3.2.2** further elucidate the concordance in genotype calls among the same individuals from these two data sets. As mentioned previously, only duplicated individuals and overlapped positions regardless of MAF were kept for this part of analysis; we ended up having 402 duplicated individuals (in 134 trios) and 7,183 overlapped positions for each data set. Based on **Figure 6** and **Table 2**, which present a histogram and detailed breakdown for comparison of genotype calls by each position across 402 duplicated pairs of individuals, there were 7,156 positions out of a total number of 7,183 positions having a mismatch count of 10 or less across all 402 duplicated pairs of individuals; in other words, 99.6% of positions have a mismatch rate of 2.5% or less. There were 9 positions having a mismatch count of 51 or higher, that is over 12.7% in terms of mismatch rate. Based on the distribution and breakdowns seen in **Figure 6** and **Table 2**, positions with over 10% mismatch rate, that is having a mismatch count of 40 or larger across all 402 duplicated pairs of individuals, were considered as outliers and thus should be removed from further analysis. In terms of comparison of genotype calls by each duplicated pair of individuals across all 7,183 positions, **Figure 7** and **Table 3** both show no significant outliers were detected; all of the duplicate pairs of individuals had mismatch counts of 141 or less across all 7,183 positions, that is less than 1.96% in terms of mismatch rate. Thus, there is no need to remove any individuals from these data. **Figures 8-9** and **Tables 4-**

**5** were generated to further examine the patterns of read depth (DP) and call quality (GQ) based on mismatch type for genotype calls in the WGS data; recall DP is a measure of coverage that quantifies the number of unique reads containing a specified nucleotide, and GQ is a measure providing a numeric value representing how certain the caller was about the call being made. Our initial assumption was that mismatches in genotype calls would occur due to lower DP or GQ, however, the findings from **Figures 8-9** and **Tables 4-5** were not consistent with this assumption. **Figure 8** and **Table 4** show the mismatched group had lower mean and median in DP compared to that of the matched group (mean: 23.9 vs. 24.9; median: 23.8 vs. 25.4), but these differences were not remarkable. On the contrary, **Figure 9** and **Table 5** illustrate the mismatched group actually had larger mean and median in GQ compared to the matched group (mean: 70.1 vs. 61.0; median: 71.6 vs. 63.6) and these differences were larger compared to the differences observed in the DP comparison. Overall, for both DP and GQ comparisons, the mismatched group had larger variabilities compared to the matched group. Therefore, our assumption was not supported by the findings based on **Figures 8-9** and **Tables 4-5**, and we conclude that mismatches in genotype calls between pairs were not a reflection of data quality in our dataset. Since we do not know which of the mismatching calls is the “true” genotype of the individual, the goal of this analysis was not to filter out any particular variant calls, but rather to look for systematic patterns in measurable variables, such as DP and GQ, that would help identify problematic positions, in the absence of duplicate samples. Based on our analysis, there is no systematic pattern to identify specific positions that should be filtered out in downstream analysis. However, the 9 positions that were considered as outliers in terms of mismatch rate can be removed from further analysis as mismatches in genotype calls even though they did not seem to have an association with poor data quality.

According to the above statements, combining the common variants from the TS and the WGS data, keeping all duplicate individuals in the WGS data and removing positions with extreme mismatch rates in genotype calls are reasonable approaches for further analysis. The findings from **Section 3.2.3** further illustrate the consistency and reproducibility of the combined data set. For this section of analysis, duplicated individuals were removed from the TS study to detect and compare the signals of linkage and association seen from two mutually exclusive groups. Additional filtering steps, as detailed in **Table 7**, were conducted at a variant level to maintain the consistency and accuracy of this study. After filtering, we ended up with having 720 individuals (in 240 trios) with 215 common variants for the TS study, and 981 individuals (in 327 trios) with 276 common variants for the WGS study. The combined data set was generated by merging the TS data and the WGS data and the same filtering procedures were

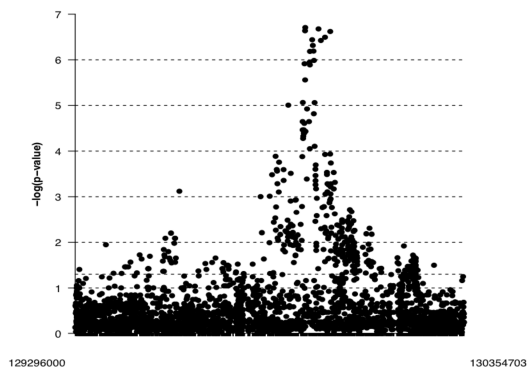
performed. After filtering, there were 1,701 individuals (in 567 trios) and 210 common variants. Signals seen in the genotypic TDT for the combined data set (**Figure 12**) were consistent with those seen in the TS and the WGS data (**Figures 10 and 11**), but were much stronger in terms of the scale of p-values. Similarly, the patterns in the LD heat map and recombination site showed good agreement among the TS, WGS and the combined data sets; two unequal-sized triangle blocks were detected with the boundary position as a genetic recombination site. Additionally, for **Figures 10-12**, there was a small “sub-block” within the large block that shows increased  $r^2$  values and the scale  $r^2$  was slightly stronger for the combined data set compared to the TS and WGS data sets. Briefly speaking, the general patterns of linkage and association were consistent among the TS, WGS and the combined data sets but the scales of signals were stronger in the combined data sets, likely due to larger sample size and concordance of genetic signal across these samples with a similar population background.

In summary, the approaches of combining the common variants in the TS data and the WGS data by keeping all duplicated individuals in the WGS data and removing positions with extreme mismatch rates in genotype calls not only showed reproducible findings consistent with previous studies, but also increased the statistical power by improving the signal intensity in genotypic TDT and LD to strengthen new findings. The work done here paves the way for using this combined data set to refine the signal in the 8q24 region. This refined signal can be used to further explore cross-population patterns to better understand the differences in signal that have been observed to date using samples from European compared to Asian populations. By narrowing down the window of genetic signal from trios of European ancestry, a more targeted approach can be taken to samples of Asian ancestry, with the goal of identifying cross-population shared genetic signal and thereby illuminate the functional role of variants in the 8q24 region on development of OFCs.

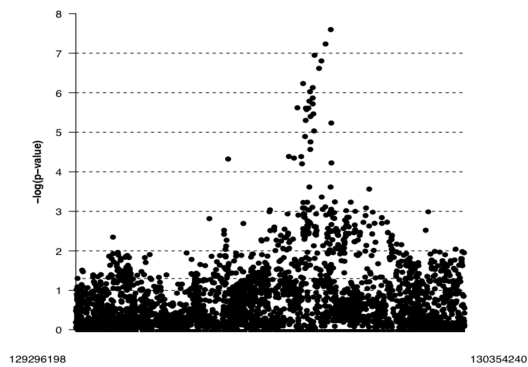
## 5 Bibliography

- Beaty, T. H., J. C. Murray, M. L. Marazita, R. G. Munger, I. Ruczinski *et al.*, 2010 A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* 42: 525-529.
- Birnbaum, S., K. U. Ludwig, H. Reutter, S. Herms, M. Steffens *et al.*, 2009 Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41: 473-477.
- Grant, K. A., C. McMahon, M. P. Austin, N. Reilly, L. Leader *et al.*, 2009 Maternal prenatal anxiety, postnatal caregiving and infants' cortisol responses to the still-face procedure. *Dev Psychobiol* 51: 625-637.
- Hartl, D. L., and A. G. Clark, 1997 *Principles of Population Genetics*. Sinauer Associates.
- Huppi, K., J. J. Pitt, B. M. Wahlberg and N. J. Caplen, 2012 The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front Genet* 3: 69.
- Laird, N. M., and C. Lange, 2006 Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet* 7: 385-394.
- Leslie, E. J., and M. L. Marazita, 2013 Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet* 163C: 246-258.
- Leslie, E. J., M. A. Taub, H. Liu, K. M. Steinberg, D. C. Koboldt *et al.*, 2015 Identification of functional variants for cleft lip with or without cleft palate in or near PAX7, FGFR2, and NOG by targeted sequencing of GWAS loci. *Am J Hum Genet* 96: 397-411.
- Li, B., W. Chen, X. Zhan, F. Busonero, S. Sanna *et al.*, 2012 A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet* 8: e1002944.
- Mossey, P., and E. Castilla, 2001 Global registry and database on craniofacial anomalies. Report of a WHO Registry Meeting on Craniofacial Anomalies.
- Murray, T., M. A. Taub, I. Ruczinski, A. F. Scott, J. B. Hetmanski *et al.*, 2012 Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between Asians and Europeans. *Genet Epidemiol* 36: 392-399.
- Spielman, R. S., and W. J. Ewens, 1996 The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59: 983-989.
- Tolarova, M. M., 2018 Pediatric Cleft Lip and Palate. Medspace.
- Uleberg, E., and T. H. Meuwissen, 2011 The complete linkage disequilibrium test: a test that points to causative mutations underlying quantitative traits. *Genet Sel Evol* 43: 20.

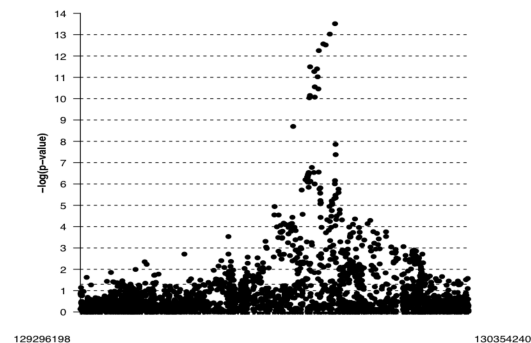
## 6 Appendix



**Figure 14.** gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the TS Study without duplicates (240 trios, 2559 common variants) for all positions on 8q24. Note that y-axis range is from 0 to 7.



**Figure 13.** gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the WGS Study without duplicates (327 trios, 3161 common variants) for all positions on 8q24. Note that y-axis range is from 0 to 8.



**Figure 15.** gTDT plot for common variants ( $MAF \geq 0.01$ ) based on the combined data without duplicates (567 trios, 2481 common variants) for all positions on 8q24. Note y-axis range is from 0 to 14.

## 7 Curriculum Vitae

### Jing (Jane) Li

• [jing.li.jli46@gmail.com](mailto:jing.li.jli46@gmail.com) • (424)343-4236 • 546 Main St, Apt 1205, New York, NY 10044

#### EDUCATION

---

##### JOHNS HOPKINS UNIVERSITY

*Master of Science in Biostatistics, GPA: 3.70/4.00*

Aug 2016 – May 2018

Baltimore, MD

##### UNIVERSITY OF CALIFORNIA LOS ANGELES

*Bachelor of Science in Statistics and Biochemistry, GPA: 3.63/4.00*

Sep 2012 – June 2016

Los Angeles, CA

#### PROFESSIONAL EXPERIENCE

---

##### ILLUMINA, INC (NASDAQ: ILMN)

*Customer Solutions Analyst Intern*

May 2017 – Aug 2017

San Diego, CA

Utilized KPIs (Key Performance Indicators) from multiple cloud based data sets to propose a new model which effectively predicted hardware failure and reduced maintenance cost by over **\$100,000**.

- **Database:** Proactively monitored the real-time performance of HiSeqX Sequencer's laser, collected data sets from Amazon Red Shift Database using SQL query, independently set up the database connection API in R environment.
- **Data Visualization & Insights:** Identified KPIs for forecasting laser failure by comparing laser performance based on failure time, created Tableau dashboard on laser performance to drive business insights.
- **Modeling:** Proposed a Decision Tree model based on potential metrics and used 10-fold cross-validation for training in order to identify malfunctioning lasers (**accuracy 90%**).
- **Public Speaking:** Presented the proposed model to over 50 people including other colleagues and managers.
- **Teamwork & A/B Testing:** Collaborated cross-functionally with the Manufacture Department to resolve instrument performance variabilities by confirming that sequencing quality differed based on changes in software version using permutation-based t test (theoretically the sequencing quality should remain the same).

##### JOHNS HOPKINS UNIVERSITY

*Research Assistant*

Apr 2017 - Present

Baltimore, MD

- **Data ETL:** Cleaned and filtered the Whole Genome Sequencing data (**over 500G**) for oral cleft family based on data quality, and conducted Mendelian Test to check for inheritance errors using vcftools on HPC.
- **Modeling & Data Visualization:** Applied familywise conditional logistic regression to examine the association between genotype transmission rate and oral cleft, created SNP plot to map the p values with genome locations.
- **Methodology:** Examined the minor allele frequency and genotype calls of duplicate positions for the WGS data and the data from previous study in preparation for merging two data sets to gain more statistical power.

##### InciteData

*Technology Analyst Intern*

July 2016 – Aug 2016

Chengdu, China

- **Machine Learning:** Cooperated with the team in conducting vehicle plate recognition algorithm using convolutional neural networks on TensorFlow. Tested run both softmax regression model and CNN model on Linux to check for the prediction precision (**accuracy 96%**).

#### PROJECTS & ACTIVITIES

---

##### DATA SCIENCE PROJECT

*Participant*

Sep 2017 – Dec 2017

Baltimore, MD

- **Web-Scraping & Shiny App Development** (<https://jhubiostatistics.shinyapps.io/mvagroup/>): Collaborated with other cohorts to predict MVA's waiting time based on location and service. Scraped the corresponding data from MVA website in real time (updated every 5 minutes) and fit distribution-based model for prediction in R. Built a Shiny app that recommends the best MVA with least waiting and driving time based on user's input.



## MIXTURE MODEL PROJECT

*Participant*

Dec 2016

Baltimore, MD

- **Algorithm Implementation:** Implemented EM algorithm, MCMC (Metropolis-Hasting) and Newton's Method from scratch in R and Python, trained the estimators to find parameters for Gaussian mixture models.

## KAGGLE COMPETITION (TOP 7)

*Participant*

May 2016

Los Angeles, CA

- **Data Processing:** Analyzed LA shelter animals' outcomes to reduce euthanasia (both training and testing data had over **110k observations**), reshaped variables and combined external data as potential predictors using R.
- **Machine Learning:** Conducted variable selections using cross validation and established final ensemble classification models using Random Forest and Boosting based on log loss function (**accuracy 94%**).

## ADDITIONAL INFORMATION

---

**Technical Skills:** R (3 yrs+)/SAS (2 yrs+)/Python/SQL/Unix/Linux/Tensorflow/STATA/SPSS/ LaTeX/Shiny App

**Database:** SQL/Amazon Red Shift Database

**Visualization:** Tableau/QGIS/Shiny App/ggplot2/plotly

**Machine Learning:** Random Forest/Boosting/EM Algorithm/Newton's Method/MCMC/SVM/Logistic Regression

**Certificate:** SAS Advanced Programmer Certificate (Mar 2016), SAS Base Programmer Certificate (Jan 2016)

**Language:** Mandarin/Japanese

**Interest:** Mountaineering/Hiking/Surfing/Snowboarding/Running/Yoga